# Statistical Analysis of the EU Data in R

**Cumali Bereket s211060 BDA SGGW**

**Female inactivity rate aged 20-64**

Project\_Cumali\_Bereket

The inactivity rate is a measure of how many people are not in the workforce. It has a big impact on the socio-economic development of a country or region, because it affects how much productive work is done. Countries and regions can try to reduce the inactivity rate by doing things like encouraging people to get back into the workforce.

# Importing the data

After searching for data on eurostat, we imported it using the code “lfst\_r\_lfp2actrtn”.

#I deal with the data : Female inactivity rate aged 20-64.And I am starting by examining how the data set is constructed.

#With the exception of geo, where there are too many values, we are able to examine the levels of each variable and factor.

#Now let’s examine the various levels (country is nut 0) and determine if they are all valid:

#We just save the countries for Nuts 0.

#We take out the countries for Nuts 2 and only keep the regions.

#Regarding the differences in the number of lines, same as with Nuts 0.

#First, we compile a list of our variables.

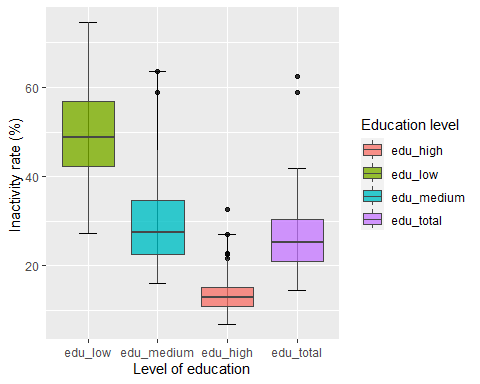
#Out of 429 values, we can see that 9 are missing for the inactivity rate. Let’s take a closer look at them:

#We quickly realize that the NRP education level contains all of the missing data. We are going to remove this modality from the remainder of the analysis because we only have nine values for it compared to 100 for the other modalities of this variable.

#For the variable of the inactivity rate, we do not observe any outliers.

#3 countries have only 8 values because they arrived or left the European Union in these 3 years. We start by making a summary of our data according to the 4 levels of education:

#In comparison to the other categories, the edu\_high category has a significantly lower average inactivity rate. On the other hand, the inactivity rate is significantly higher in the edu\_low category.



#The boxplots above show what we were talking about before: the edu\_high category has low inactivity, edu\_low has high inactivity, edu\_medium and edu\_total are about the same level.

#First of all we make a summary of our variables.

#On 3520 values, we can see that there are 252 missing values for the inactivity rate. Let’s look at them in detail:

#Contrary to before, missing values are more distributed here, whether it is for the education level, the region or even the year.

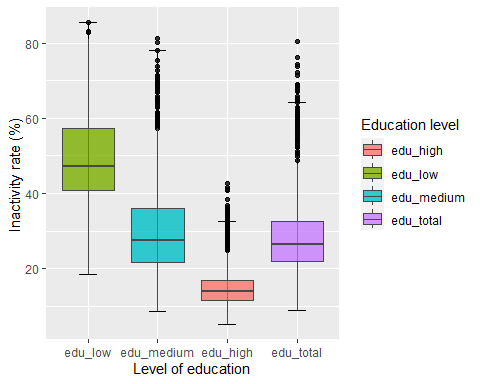
#As with Nuts 0, we will remove NRP in a first step

#When we examine our data, we have a single case where the inactivity rate is equal to 0 for the ES63 region in 2020 and for the high education category. We decide to keep the value anyway.

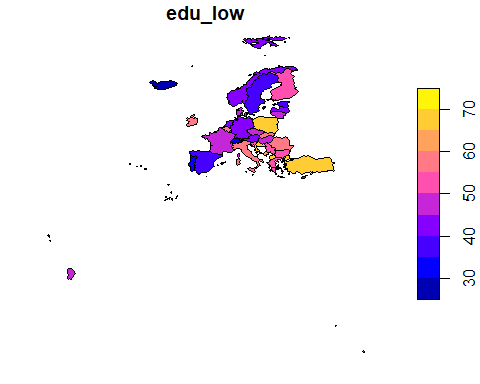
#The same as in the previous case, the regions that are not equal to 12 come from countries that have arrived or left the European Union in these 3 years.

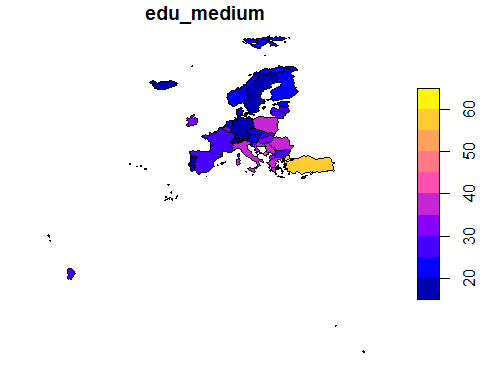
#We start by making a summary of our data according to the 4 levels of education:

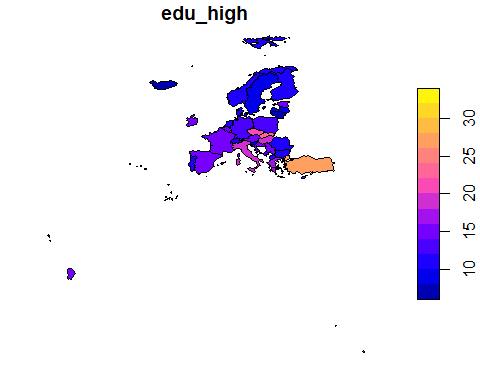
#In comparison to the other categories, the edu\_high category has a significantly lower average inactivity rate. On the other hand, the inactivity rate is significantly higher in the edu\_low category. Similar to NUTS 0, exactly.

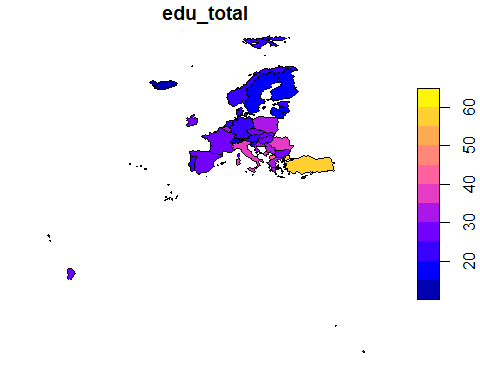


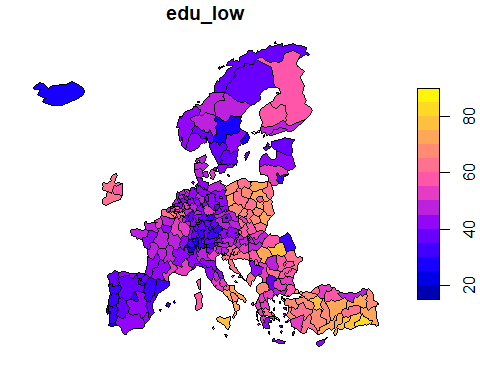
The boxplots above show what we were talking about before: the edu\_high category has low inactivity, edu\_low has high inactivity and edu\_medium and edu\_total are about the same level. Same as with Nuts 0.

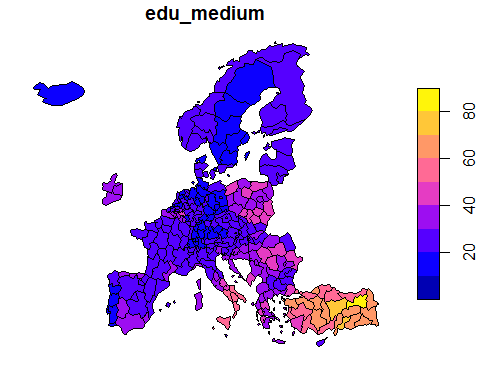


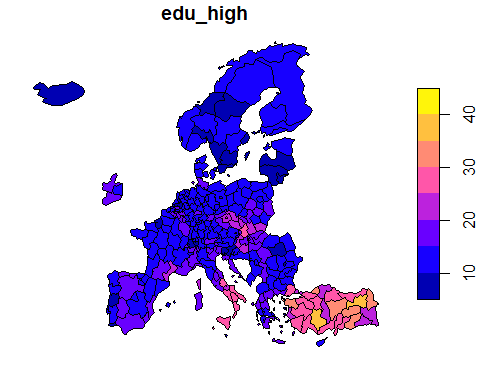


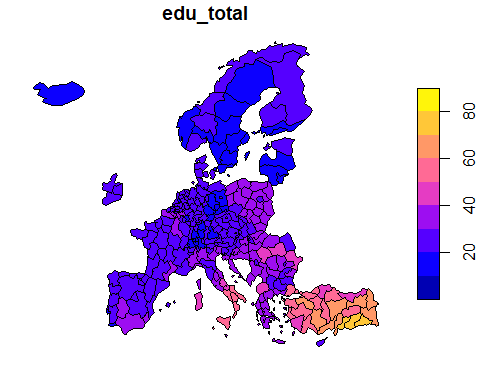












We begin by categorizing education levels. We only need to get rid of the mode “edu total” to accomplish this.

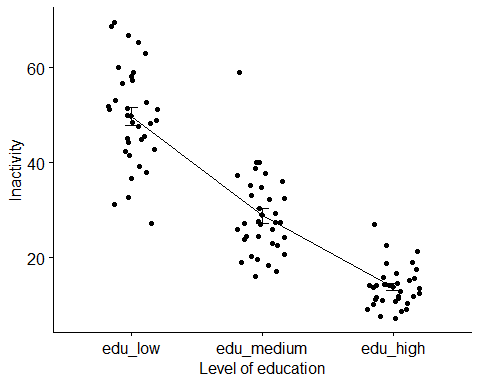
#We test the normality of the distribution with the Shapiro test with the following hypothesis

#H0 : The data come from a normal distribution H1 : The data do not come from a normal distribution

#The Shapiro test confirms this with a pvalue < alpha(5%).

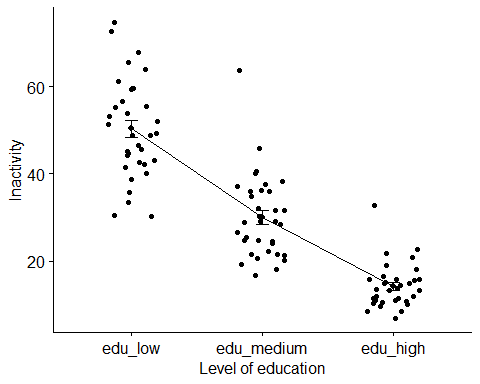
#We cannot say that the sample follows a normal distribution, so we will therefore use a non-parametric Anova.

#We can visualize our data.



#2 of the 3 groups has a pvalue < 0.05, so we reject h0. We cannot say that the sample follows a normal distribution, so we will therefore use a non-parametric Anova.

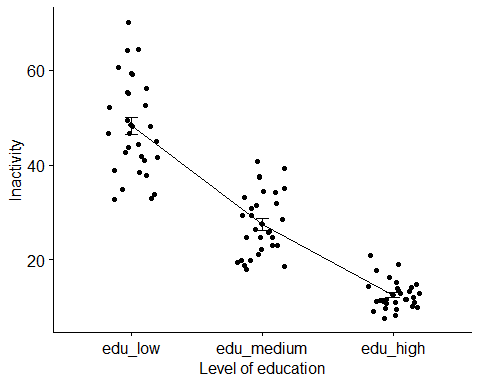
#We can visualize our data.



#The 3 groups have a pvalue > 0.05, so we cannot reject h0.

#We can say that the sample follows a normal distribution, so we will therefore use a parametric Anova.

#We can visualize our data.



#We now reverse the analysis, i.e. we will analyze whether the year has an impact on the inactivity rate, all this for the 3 education levels.

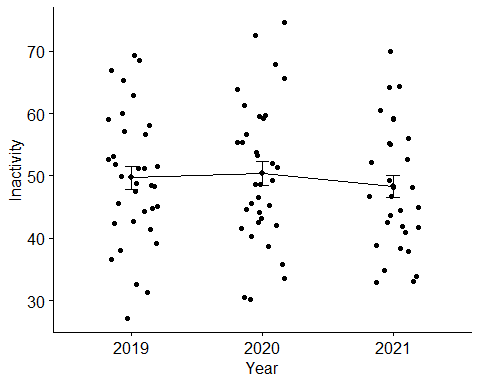
#We will remove the modality “edu total” like previously because there is no interest in having it.

#We then divide our dataset in 3, for each level of education.

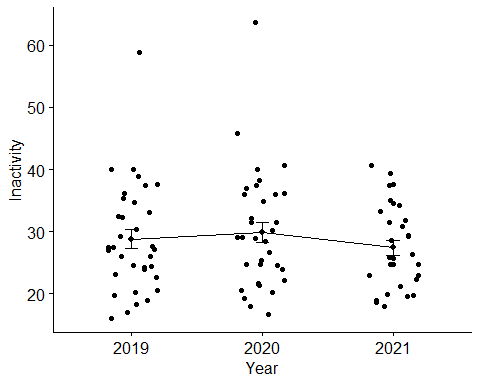
#After this is completed, we can begin our analysis for each educational level.

# #Low

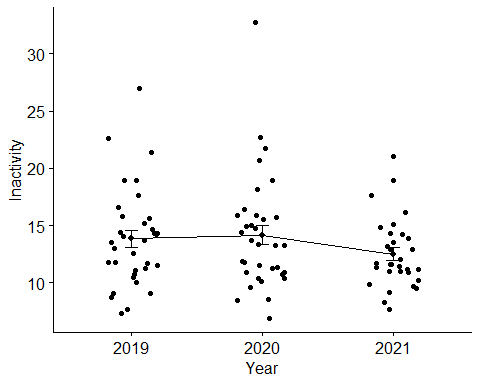
#To get started with our analysis, we can look for normality.



To get started with our analysis, we can look for normality.

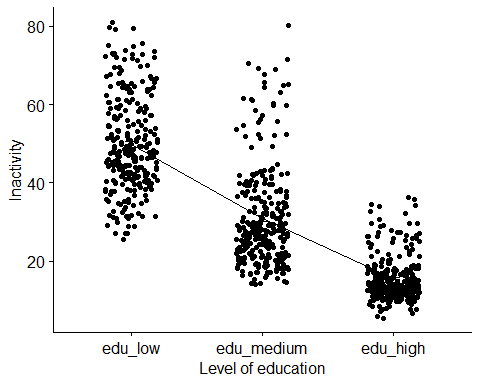


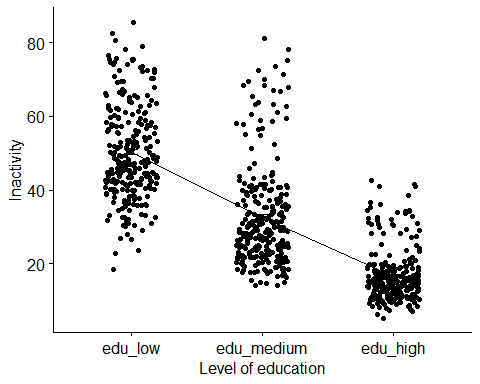
#To get started with our analysis, we can look for normality.



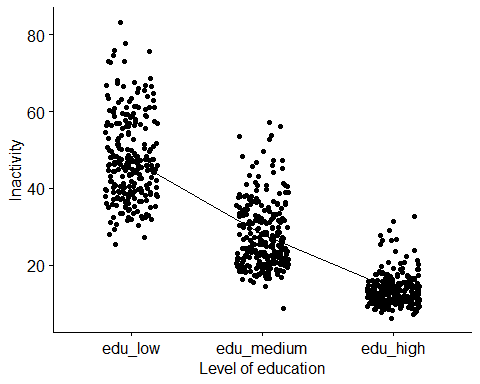
#Once this is finished, we can begin each year’s analysis. To get started with our analysis, we can look for normality.

#All pvalues are < 0,05, so we reject H0. We will use a non-parametric Anova because we cannot say that the sample follows a normal distribution. We can visualize our data.





#The 3 groups have a pvalue < 0.05, so we can reject h0. We will use a non-parametric Anova because we cannot say that the sample follows a normal distribution. We can visualize our data.

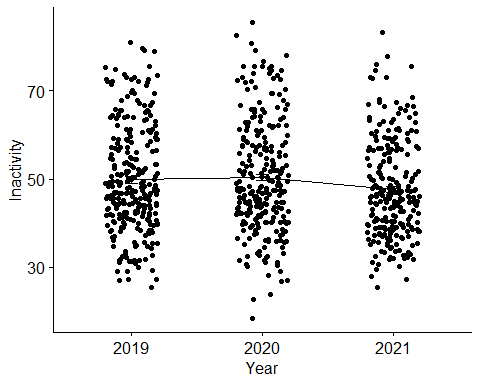


#We now return to the analysis of the years. #Since there is no interest in having the modality “edu total,” we will remove it as previously.

#We then divide our dataset in 3, for each level of education.

#After this is completed, we can begin our analysis for each educational level.

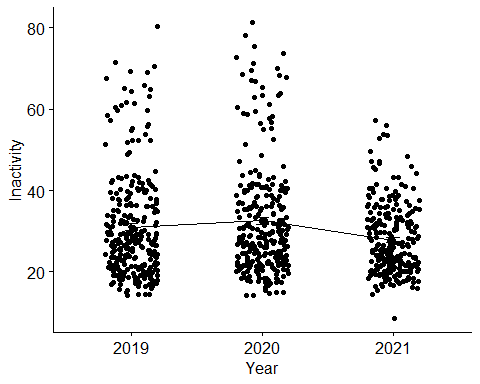
#To get started with our analysis, we can look for normality.



#To get started with our analysis, we can look for normality.

#Here, all Pvalues are < 0,05 so we can reject h0.

#We will use a non-parametric Anova because we cannot say that the sample follows a normal distribution. We can visualize our data.



#To get started with our analysis, we can look for normality.#Here, all Pvalues are < 0,05 so we can reject h0. We will use a non-parametric Anova because we cannot say that the sample follows a normal distribution. We can visualize our data.

